1.1 **Prix de billets de TGV espagnols**: Les données `renfe` contiennent des informations sur 10 000 billets de trains vendus par la compagnie Renfe, l'entreprise ferroviaire publique espagnole. Les données incluent les variables:

- `prix`: prix du billet (en euros);
- `dest`: indicateur binaire du trajet, soit de Barcelone vers Madrid (`0`) ou de Madrid vers Barcelone (`1`);
- `tarif`: variable catégorielle indiquant le tarif du billet, un parmi `AdultoIda`, `Promo` et `Flexible`;
- `classe`: classe du billet, soit `Preferente`, `Turista`, `TuristaPlus` ou `TuristaSolo`;
- `type`: variable catégorielle indiquant le type de train, soit Alta Velocidad Española (`AVE`), soit Alta Velocidad Española conjointement avec TGV (un partenariat entre la SNCF et Renfe pour les trains à destination ou en provenance de Toulouse) `AVE-TGV`, soit les trains régionaux `REXPRESS`; seuls les trains étiquetés `AVE` ou `AVE-TGV` sont des trains à grande vitesse.
- `duree`: longueur annoncée du trajet (en minutes);
- `jour` entier indiquant le jour de la semaine du départ allant de dimanche (`1`) à samedi (`7`).

On considère le temps de parcours pour les trains à grande vitesse (`AVE` et `AVE-TGV`). Le temps médian entre les deux villes dans la « population » est de $\nu = 2.833$ heures, tandis que la moyenne de la « population » est de $\mu = 2.845$ heures; ces quantités ont été déterminées sur la base des données complètes contenant plus de 2.3 millions d'entrées et sont donc considérées comme connues, contrairement à la plupart des applications pratiques.
Une étude de simulation a été conduite pour déterminer le comportement de tests pour un échantillon. L'algorithme suivant a été répété 10 000 fois:

(a) sélection d'un sous-échantillon de taille $n = 100$.
(b) calcul de la statistique du test-t pour un échantillon correspondant à $\mathscr{H}_0 : \mu = \mu_0$ (versus $\mathscr{H}_0 : \mu \neq \mu_0$) pour différentes valeurs de μ_0.
(c) calcul de la statistique du test des signes pour le test bilatéral $\mathscr{H}_0 : \nu = \nu_0$ pour différentes valeurs de ν_0.
(d) calcul de la statistique du test des rangs signés de Wilcoxon pour le test bilatéral $\mathscr{H}_0 : \nu = \nu_0$ pour différentes valeurs de ν_0.
(e) sauvegarde des valeurs-p associées à chacun des trois tests.

Notez que le test des signes et le test de Wilcoxon sont deux tests pour la **médiane**.
La fig. 1 montre le pourcentage de valeur-p parmi les 10 000 qui sont plus petites que 0,05, c'est-à-dire la proportion de rejet (à un niveau de 5%) de $\mathscr{H}_0 : \mu = \mu_0$ contre l'alternative bilatérale à $\mu_0 \in \{2,83; \mu; 2,835; 2,84; \ldots; 2,995; 3\}$ (pour le test des signes et de Wilcoxon, nous testons si la médiane est égale à ces même valeurs). Utilisez la courbe de puissance (Figure 1) pour les trois tests de localisation afin de répondre aux questions suivantes:

(a) Expliquez pourquoi la proportion de rejet de chaque test augmente quand on se déplace vers la droite sur le graphique.
(b) Supposez que l'on répète l'expérience de simulation, mais cette fois avec des sous-échantillons aléatoires de taille $n = 1000$. Comment est-ce que les points pour le test-t pour un échantillon se compareraient à ceux tracés sur le graphique? Seraient-ils en dessous, à la même hauteur ou au dessus?
(c) Expliquez pourquoi la valeur sur le graphique pour le test-t pour un échantillon **devrait être** approximativement 0,05 dans un voisinage de $\mu = 2,845$ (idem pour le test des signes et le test de Wilcoxon, où les valeurs devraient être approximativement 0,05 autour de $\nu = 2,833$).
(d) Selon la Figure 1, à quelle fréquence rejeteriez-vous l'hypothèse nulle pour le test des rangs signés de Wilcoxon à $\nu = 2,833$? Expliquez les conséquences de cette trouvaille sur votre inférence.
(e) Produisez un diagramme quantile-quantile normal et commentez sur la robustesse du test-t à des déviations de l'hypothèse de normalité.

1.2 Supposez que l'on veut comparer le tarif moyen pour les trains à grande vitesse pour les deux destinations, soit de Madrid vers Barcelone et le trajet inverse de Barcelone à Madrid. Une étude de simulation a été réalisée dans laquelle le test de Welch pour deux échantillons a été calculé sur des sous-échantillons aléatoires de taille $n = 1000$. Les données `renfe_simu` contiennent les différences moyennes (`difmoy`), les statistiques de test (`Wstat`), les

Compilé le 14/12/2020 à 15:50

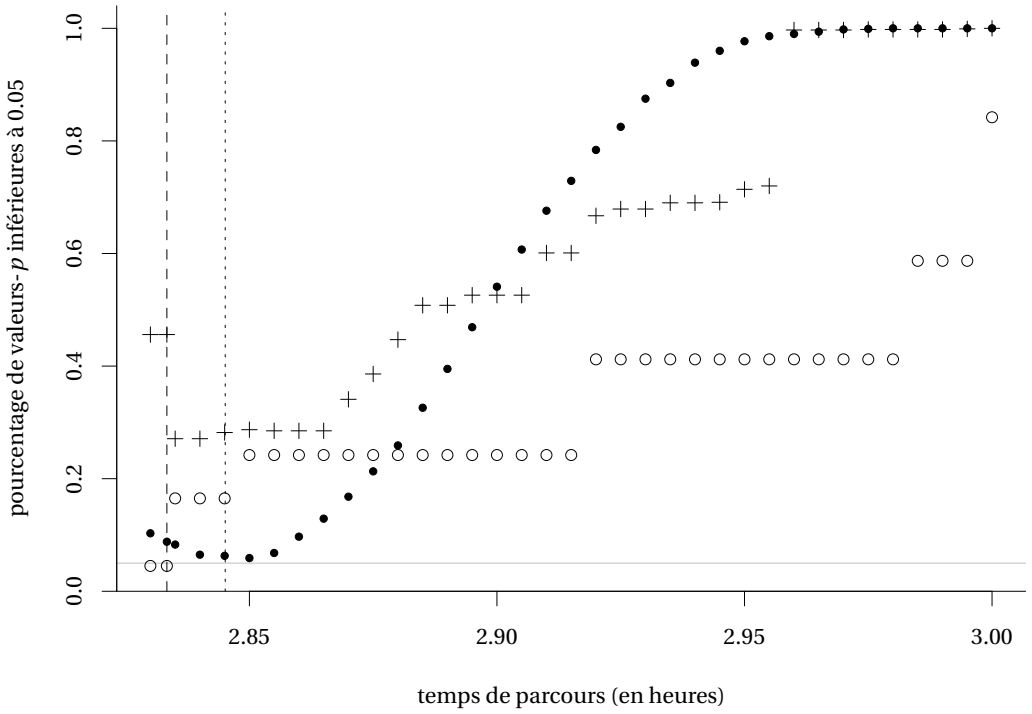

Figure 1: Courbe de puissance pour trois tests de localisation, soit le test-t pour un échantillon (disque), le test des rangs signés de Wilcoxon (croix) et le test des signes (cercles), en fonction du temps de parcours (en heures). La ligne horizontale grise correspond à 0,05, tandis que la ligne traitillée verticale indique la vraie médiane ν et la ligne pointillée verticale marque la vraie moyenne μ.

valeurs-p (`valp`) et les intervalles de confiance à 95% (`icbi` et `icbs`) pour 1000 répétitions. Supposez que l'on sait que la vraie différence moyenne dans la population est de $-0,28$€. Utilisez les données simulées pour répondre aux questions suivantes et **commentez brièvement** sur chaque sous-question.

(a) Quel est le taux de couverture empirique des intervalles de confiance à 95% (c'est-à-dire le pourcentage des intervalles couvrant la valeur de la « vraie » différence moyenne)?

(b) Tracez un histogramme des différences moyennes et superposez la vraie différence moyenne à l'aide d'un trait vertical.

(c) Calculez la puissance du test (pourcentage de rejet de l'hypothèse nulle sous l'hypothèse alternative).

1.3 À l'aide des données `renfe`, testez si le prix moyen du billet pour un train de classe `AVE-TGV` est le même que celui d'un train régio-express (`REXPRESS`). Veillez à

- énoncer l'hypothèse nulle et l'hypothèse alternative,
- justifier avec soin le choix de votre statistique de test,
- rapporter la différence moyenne estimée et un intervalle à 90% pour cette différence,
- conclure dans le cadre de la mise en situation.

1.4 Les données fictives `assurance` font partie du livre "Machine Learning with *R*" de Brett Lantz (2003). La base de donnée contient des informations sur les frais médicaux facturés à 1338 adultes américains assurés au courant de l'année 2003, soit

- `age`: âge (en années)
- `sexe`: sexe, `homme` ou `femme`,
- `imc`: indice de masse corporelle (en kg/m^2),
- `enfant`: nombre d'enfants à charge,
- `fumeur`: `oui` pour les fumeurs, `non` autrement,
- `region`: lieu de résidence, une région parmi `sudouest`, `sudest`, `nordouest` ou `nordest`,
- `frais`: les frais médicaux annuels (en dollars USD).

Selon l'Organisation mondiale de la Santé (OMS), l'indice de masse corporelle (IMC) permet de classer les individus conformément à une échelle allant de l'insuffisance pondérale à l'obésité morbide (classe III). Ladite classification est définie dans le tableau Table 1.

Classification	IMC (kg/m^2)
< 18.5	Insuffisance pondérale
18.5–24.9	Corpulence normale
25.0–29.9	Surpoids
30.0–34.9	Obésité
35.0–39.9	Obésité de classe II et III

Table 1: Classification internationale de l'obésité chez les adultes selon l'OMS.

À l'aide des données `assurance`, répondez aux questions suivantes.

(a) Effectuez une analyse exploratoire des données: quelles sont les aspects les plus importants pour expliquer les frais médicaux?

(b) Les fumeurs paient-ils des frais médicaux en moyenne équivalents aux non-fumeurs? Justifiez adéquatement votre réponse

(c) Les fumeurs considérés obèses (`imc` ≥ 30) paient-ils des frais médicaux en moyenne plus élevés que les fumeurs non obèses? Donnez trois intervalles de confiance à 90%, 95% et 99% pour estimer la différence moyenne dans ce contexte. Comparez les intervalles et expliquez les différences observées selon le niveau.

1.5 **Calcul de la puissance d'un test statistique**

Le programme SAS `puissance.sas`, écrit par Rick Wicklin de SAS Institute Inc., contient du code pour faire

une étude de simulation afin de calculer la puissance d'un test-t pour deux échantillons (l'équivalent d'un modèle linéaire simple avec une variable binaire comme variable explicative).

(a) Expliquez brièvement dans vos mots en quoi consiste les étapes de l'étude de simulation.

(b) Tracez le graphique pour les paramètres $n_1 = n_2 = 10$, $\sigma = 1$ et $B = 10\,000$ et commentez sur l'apparence de ce dernier.

(c) Faites varier le nombre de simulations de $B = 100$ à $B = 10\,000$. Que remarquez-vous quand le nombre de simulation est petit? Expliquez pourquoi cet effet disparaît quand la nombre de simulations augmente.

(d) Modifiez le code pour que la taille des groupes soit $n_1 = 10, n_2 = 30$ et $n_1 = 20, n_2 = 20$. Dans lequel des deux scénarios la puissance est-elle plus élevée et pourquoi?

(e) Modifiez le code pour simuler $n = 20$ observations dans chaque groupe de lois normales d'écart-type $\sigma_1 = 1$ et $\sigma_2 = 5$ avec $B = 100\,000$. Rapportez l'estimé de l'erreur de type I avec le nombre de simulations et un intervalle de confiance ponctuel à 90% pour l'erreur de type I; expliquez comment vous avez dérivé ce dernier.

Compilé le 14/12/2020 à 15:50